Néologismes, dictionnaires et informatique

Emmanuel Cartier, Jean-François Sablayrolles

HAL JFS 29 2009a

Sablayrolles Jean-François (Paris 13 et LDI UMR 7187) et Cartier Emmanuel (Paris 13 et LDI UMR 7187)

« Néologismes, dictionnaires et informatique »

Cahiers de lexicologie n° 93, 2008-2, p. 175-192. (février 2009)

Résumé

La veille néologique se fonde généralement sur un ou des dictionnaires pris comme corpus d'exclusion. Ce qui pose des problèmes tant théoriques que pratiques. Le développement de l'informatique change en partie la donne. D'une part, des dictionnaires électroniques récents à la nomenclature plus abondante et intégrant les formes fléchies, comme Morfetik, sont plus efficaces pour la détection des néologismes formels (y compris flexionnels). D'autre part, les dictionnaires électroniques fondés sur les classes d'objets permettent de détecter automatiquement des types de néologismes qui échappaient à l'extraction automatique classique : néologismes syntaxiques, sémantico-syntaxiques et néologismes formels homonymiques de mots conventionnels. Le choix des documents (papier, informatisés ou uniquement informatiques) soumis à l'extraction de néologismes dépend des objectifs que l'on se donne : lexicographie et étude de l'évolution de la langue ou étude de la créativité lexicale des locuteurs. Les moteurs de recherches permettent de mesurer la circulation des néologismes et d'étudier plus précisément tel ou tel du fait de son intérêt supposé. Les néologismes incorporés sont intégrés dans une base de données, Neologia, construite conjointement par des linguistes et des informaticiens en fonction des analyses que l'on souhaite pouvoir conduire, à l'aide de requêtes simples (proportion de néologismes par catégories grammaticales, par matrices lexicales, etc.) et de requêtes croisant plusieurs critères afin de mesurer les variations concomitantes des types d'énonciateurs, des types de néologismes et des situations d'énonciation. Les spécificités de la néologie et de la créativité lexicale du français hexagonal contemporain apparaîtront ainsi plus clairement. La base Neologia est ainsi présentée dans son fonctionnement et dans deux de ses interactions majeures avec l'extérieur : l'incorporation automatique de néologismes et l'enrichissement des dictionnaires généraux du laboratoire LDI.

Mots clés : Néologie, base de données, traitement informatique, lexicographie et métalexicographie

NEOLOGISMES, DICTIONNAIRES ET INFORMATIQUE

Toute langue vivante est en constante évolution et une des marques les plus saillantes de cette instabilité est le renouvellement de son lexique, avec la disparition de mots obsolètes (peu étudiée) et l'apparition de nouvelles unités : les néologismes. L'étude de ceux-ci est d'autant plus incontournable dans l'étude d'une langue qu'ils touchent à tous les aspects de la langue, partant à toutes les branches des sciences du langage : phonologie, morphologie, syntaxe, sémantique, énonciation, sociolinguistique, pragmatique… La lexicographie et la métalexicographie sont également concernées, mais les relations entre néologismes et dictionnaires changent avec l'avènement de l'outil informatique. Le choix des corpus d'exclusion et des corpus d'étude pour la veille néologique ainsi que les types d'études conduites sur les données recueillies ont évolué avec les possibilités offertes par l'informatique. Cet article présente les interactions entre la néologie et l'informatique sous deux points de vue principaux : d'une part, vis-à-vis du repérage des néologismes dans les corpus, car l'outil informatique, de ce point de vue, offre des possibilités de traitement bien plus grandes et systématiques, qui permettent d'envisager une veille néologique efficace ; d'autre part, vis-à-vis du stockage des néologismes, car les bases de données offrent une structure d'accueil aujourd'hui incontournable.

1. Néologismes et dictionnaires comme corpus d'exclusion

L'absence d'une entrée dans un dictionnaire vaut traditionnellement certificat de néologicité pour une unité lexicale soupçonnée de nouveauté. Mais ce recours aux dictionnaires (papier ou informatisés) comme corpus d'exclusion pose de nombreux problèmes que l'apparition de nouveaux outils ou des conceptions différentes des dictionnaires seront à terme en mesure de résoudre.

1.1. Les dictionnaires papier et leurs versions informatisées

Les problèmes posés par les dictionnaires comme corpus d'exclusion des néologismes sont bien connus et ont fait l'objet d'études antérieures (voir, entre autres, Sablayrolles 2000, 2002 et 2008). Contentons-nous donc d'en rappeler quelques-uns.

Les différences dans leur nomenclature constituent la première source de difficultés : quel(s) dictionnaire(s) privilégier : de langue ou encyclopédique, monovolumaire ou plurivolumaire, généraux ou de spécialité, etc. ? Faut-il s'aligner sur le plus disant, le moins disant, exiger, pour nier le caractère néologique de l'item en question, la présence dans la majorité des dictionnaires consultés ou du moins l'accord de plusieurs dictionnaires, voire la présence dans un seul malgré l'adage *Testis unus, testis nullus* « un seul témoin, pas de témoin » ? Aucune réponse vraiment satisfaisante ne pouvant être apportée, on se contente de pratiques plutôt aléatoires.

Par ailleurs l'existence dans un/des dictionnaire(s) d'une entrée ayant la même forme que l'item testé aboutit à exclure systématiquement aussi bien les faits de néologie sémantique (*détricoter* des opérations informatiques comme un tricot) que les néologismes formels par homonymie (*biofilm* terme traditionnel de biologie « ensemble des bactéries qui colonisent un milieu » et néologisme dénommant un genre cinématographique, équivalent de *biopic*, « film retraçant la vie d'une personne connue »).

Enfin la lourdeur de la consultation de dictionnaires papier est avérée et constitue un obstacle pour la veille néologique : elle prend beaucoup de temps et d'énergie, exige d'avoir un grand nombre de documents à disposition et donc limite les lieux où le travail de vérification peut être effectué.

Les versions informatisées des dictionnaires papier lèvent ce dernier inconvénient en laissant les autres intacts et en y ajoutant d'autres : tous les dictionnaires ne possèdent pas de telles versions informatisées, et ceux qui en possèdent ne sont pas actualisés chaque année. Malgré leur variabilité, leur incomplétude (revendiquée : aucun dictionnaire ne prétend tout enregistrer), leur frilosité qui fait parfois différer des décennies l'entrée de mots qui circulent couramment[1], les dictionnaires constituent des outils irremplaçables dans la mesure où ils suppléent les ignorances lexicales que tout locuteur ne manque pas d'avoir dans sa propre langue et où ils sont le résultat de travaux d'équipe s'appuyant sur une longue tradition, ce qui tend à leur constante amélioration. On ne peut donc pas s'en passer, mais on ne peut les utiliser mécaniquement et on doit faire preuve d'une grande vigilance à leur égard. Leur manque de systématicité dans leur description constitue une pierre d'achoppement que deux types d'outils construits au laboratoire LDI ont vocation à faire disparaître, rendant l'extraction (semi)-automatique des néologismes plus fiable et efficace.

1.2. Le dictionnaire électronique Morfetik

Élaboré depuis plusieurs années par Michel Mathieu-Colas (voir Mathieu-Colas, 2006) qui a collationné des données prises à diverses sources existantes, le dictionnaire électronique Morfetik comprend plus de 750 000 entrées d'unités monolexicales auxquelles s'ajouteront bientôt quelque 250 000 formes d'unités polylexicales (dès que les vérifications sur la forme graphique des pluriels des mots composés auront été achevées). Une des spécificités de ce dictionnaire consiste à entrer les formes flexionnelles, de genre et de nombre ainsi que de la conjugaison verbale, avec indication des doublets et des formes rares. Ce sont ainsi plusieurs centaines de milliers de formes que contient Morfetik. Ce dictionnaire constitue un corpus d'exclusion idéal pour l'extraction automatique des néologismes formels (y compris les néologismes flexionnels comme *ils closirent* du verbe *clore* dont Littré regrette explicitement la défectivité) et un travail d'une étudiante est en cours pour son exploitation avec cet objectif. Les éléments qui seront signalés comme inconnus seront soit des « fautes » diverses (de frappe, lapsus, etc.), soit des lacunes accidentelles (toujours possibles) du dictionnaire, qui devront alors être comblées, soit des néologismes. Parmi ceux-ci, qui intéressent tous l'équipe néologie et qui seront tous entrés dans la base de néologismes Neologia, il faudra, en fonction de leur circulation observable, décider ceux qui devront être

[1] Ainsi l'expression familière *à toute blinde* « vite » datée de 1931 par le *Petit Robert* (*PR*) est encore absente du *Petit Larousse Illustré* (*PLI*) 1991, soixante ans après, mais elle figure dans l'édition de 2008. *Hype* daté des années 80 par *PR* est entré dans ce dictionnaire entre 1996 et 2008, mais il était toujours absent du *PLI* 2008 et ne figure pas dans la liste des mots introduits en 2009. Le sigle *CLUF* (*Contrat Licence Utilisateur Final* pour les jeux électroniques diffusés par l'internet), pour lequel les moteurs de recherche indiquent des dizaines de milliers de sites, est encore absent du *Nouveau Littré* 2006, du *Petit Robert* 2007 et du *Petit Larousse* 2008 etc. On pourrait multiplier les exemples à l'infini.

entrés dans le dictionnaire Morfetik et ceux qui, encore hapax ou de très faible diffusion, ne le seront pas et ne resteront consignés que dans la base Neologia (présentée en 3).

1.3. Dictionnaire électronique de phrases élémentaires

Une propriété fondamentale des autres dictionnaires établis au laboratoire LDI avec le modèle des classes d'objets réside dans le fait qu'ils sont des dictionnaires de phrases élémentaires et non des dictionnaires de mots (v. Gross 1994 entre autres). Une phrase élémentaire étant un prédicat saturé par ses arguments, il faut noter pour chaque classe de prédicats les arguments qu'il appelle et pour les arguments de quel prédicat ils peuvent être des actants. Les dictionnaires électroniques de phrases élémentaires font de manière systématique et délibérée ce que les dictionnaires traditionnels font de manière non avouée et non systématique, avec une explicitation des schémas syntactico-sémantiques des prédicats, indiquant les constructions syntaxiques et les restrictions sémantiques (en termes de classes d'objets) des arguments (voir Grezka et al., 2007, et Cartier, 2007, pour l'utilisation de ces schémas). L'utilisation de tels dictionnaires comme corpus d'exclusion ouvre la voie à la détection automatique de nouveaux emplois de signifiants déjà attestés et va bien au-delà d'un sous-ensemble de la néologie formelle (les nouveaux signifiants non homonymes d'items déjà existants) que permettaient de relever les nomenclatures des dictionnaires traditionnels pris comme corpus d'exclusion : le saut n'est pas seulement quantitatif, il est aussi et surtout qualitatif. La mise en place de ces systèmes d'extraction automatique prendra du temps, mais certaines expériences conduites par une doctorante du laboratoire sur les adjectifs relationnels donnent des résultats prometteurs. À terme ce sont trois grands groupes de néologismes supplémentaires qui pourront être détectés automatiquement.

Des changements de construction syntaxique sans changement sémantique notable seront repérés. Cette catégorie de la néologie syntaxique est méconnue bien qu'elle figure déjà dans le *Grand Dictionnaire Universel* de Pierre Larousse dans la partie encyclopédique de l'article *néologie*. Il s'agit essentiellement de changements dans le schéma argumental avec des constructions directes remplaçant des constructions indirectes ou vice-versa (comme le remplacement de plus en plus fréquent de *vitupérer qqun* par *vitupérer contre qqun*) ou les échanges entre constructions transitives et intransitives (*ironiser qqun* ou *ironiser un passage*, formulations raccourcies de *parler de qqun de manière ironique* ou *conférer un tour ironique à un énoncé*), des constructions directes au lieu de constructions factitives (*signer un artiste*, « faire signer un contrat à un artiste » pour une maison de disques)…

Beaucoup plus fréquemment ce sont des innovations à la fois syntaxiques et sémantiques qui pourront être relevées, avec des emplois polysémiques nouveaux qui se traduisent par des changements dans le schéma argumental avec des changements de sens (extension, restriction de sens et emplois tropologiques de type métaphoriques, métonymiques, etc.). Ainsi le verbe *détricoter* a vu s'élargir progressivement son domaine d'emploi originellement restreint à la classe des objets concrets <objet manufacturé, tricot> comme second argument (COD). Une première extension a fait employer ce verbe pour des textes comme celui du projet de Constitution européenne que son auteur ne voulait pas voir détricoter. Avec le sens plus général de « défaire, détruire », une deuxième extension a fait appliquer ce prédicat verbal au contenu de textes : les syndicats et certaines personnalités politiques accusaient le gouvernement de vouloir détricoter le code du travail ou même les acquis sociaux dont il était le garant. Une autre extension avec le sens de « défaire dans le sens exactement inverse de ce que l'on a fait » est attestée dans le domaine informatique :

> Comment rattraper une erreur avec VigiPaiement : Quatrième erreur : il faut détricoter. Vous vous êtes entêté(e) dans l'erreur en faisant une facture fictive, payée par un chèque introuvable, remis en banque sans vérifier, mais, heureusement, pas encore rapproché. Et justement voici l'extrait bancaire qui vous alerte. Vous n'avez pas le choix, il faut tout détricoter, dans l'ordre inverse. Retrouver la remise de chèques dans l'onglet Recettes, faire un double clic pour l'afficher, retirer le chèque introuvable et enregistrer la remise corrigée. Retrouver le bordereau dans l'onglet Paiements pour le supprimer. Retrouver la fausse facture dans l'onglet Factures, la sélectionner, maintenir la touche ctrl en cliquant dans la liste et relâcher sur Supprimer » (sur le site internet prokov.com, sans indication de date).

Ce sont enfin des innovations à la fois formelles et sémantiques qui peuvent être aussi détectées, même en l'absence d'une différence formelle avec un ou d'autres items déjà existants dans le lexique. Des créations homonymiques, avec des schémas morphologiques différents, présentent des sens différents. C'est ce que montre un article de Danielle Corbin (1990) au titre éloquent « Homonymie structurelle et définition des mots construits, vers un dictionnaire dérivationnel » avec l'analyse de *antialcoolique* comme illustration : quatorze homonymes possibles dont quatre sont attestés. Ainsi *endormissement* au sens de « état de quelqu'un qui est mal réveillé, pas vif » est-il homonymique du conventionnel *endormissement* « passage de la veille au sommeil ». Dans ce cas comme dans d'autres similaires, on ne doit pas parler de néologie sémantique comme on l'entend

souvent, car les deux emplois ne dérivent pas l'un de l'autre (de quel lien sémantique pourrait-on naturellement faire l'hypothèse pour passer du sens conventionnel au sens du néologisme ?) mais ils sont créés indépendamment, par des règles différentes et/ou sur des bases différentes. En l'occurrence la lexie conventionnelle a comme base l'infinitif *s'endormir* (c'est « le fait de s'endormir ») alors que le second, néologique, a comme base le participe passé *endormi* au sens de « pas vif » (« le fait d'être endormi, pas vif »). Dans cette direction, est à l'étude —avec une collaboration de linguistes et d'informaticiens du laboratoire— un projet d'analyseur automatique reconnaissant les formants (morphèmes et segments morphologiques, voir Touratier 2002) d'un néologisme et, sous forme d'algorithme, la ou les structure(s) morphologique(s) qui peuvent y être associées (les différentes possibilités de création, avec un sens pour chacune des structures reconnues). *Déprécarisation*, analysable en quatre morphèmes (préfixe *dé-*, radical *précar-* allomorphe de *précaire*, suffixe *-is* formant des verbes et suffixe *-ation* formant des noms) peut être la préfixation du nom *précarisation* « le contraire de la précarisation, situation de non précarisation » ou la suffixation du verbe possible non attesté *déprécariser*. Le contexte « depuis 2003 une politique de déprécarisation a été menée » (un responsable de la Poste, *Métro*, 21 mars 2008) invite à préférer cette deuxième solution : il s'agit de l'action de « déprécariser », de mettre fin à des situations précaires.

Pour conclure sommairement cette première partie, on peut dire que les conceptions développées au laboratoire LDI avec le modèle des classes d'objets et certains outils informatiques qui y sont élaborés permettent de résoudre des problèmes théoriques relatifs au statut de la néologie : par rapport à quoi mesurer la nouveauté (c'est une des questions fondamentales qu'avait identifiées A. Rey 1976). C'est en effet moins la néologie qui pose des difficultés de définition que les insuffisances des descriptions du fonctionnement de la langue et du lexique en particulier. On manque d'outils fiables qui servent de pierre de touche pour apprécier le caractère innovant d'un emploi dans un énoncé. Mais de la théorie à la mise en pratique, il y a tout un travail qui ne pourra se développer que pas à pas.

2. Constitution de corpus et veille néologique

2.1. Dépouillement manuel de corpus écrits et néologismes oraux

Bien qu'il soit dévoreur de temps, qu'il limite le nombre et la dimension des corpus traités et qu'il soit sujet à des inattentions et oublis, le dépouillement manuel traditionnel a encore de beaux jours devant lui. Dans l'attente de la mise au point et de l'utilisation effective à grande échelle des corpus d'exclusion évoqués ci-dessus, nombre de néologismes syntaxiques, sémantiques et homonymiques ainsi que des néologismes oraux ne sont encore repérables que manuellement, sans compter qu'il serait dommage de ne pas retenir des néologismes rencontrés au hasard des lectures ou d'écoute en dehors des corpus exploités systématiquement.

Bien sûr ces relevés sont sujets à des distorsions d'incorporation entre collecteurs (Gardin *et alii* 1974) et même à des intermittences du sentiment néologique chez un même collecteur au cours du temps. Mais ces difficultés ne sont pas propres aux relevés manuels ni à l'étude de la néologie. Des expériences sur les variations du sentiment néologique (Sablayrolles 2003 et Benhariz à paraître en 2009) ont permis d'en mettre à jour certains facteurs (âge, formation initiale, centres d'intérêt, stratégies…) afin de les réduire en explicitant le plus clairement possible les critères de néologicité. Un des intérêts de ces incorporations manuelles est précisément la confrontation des relevés qui permet d'affiner le concept de néologie et du coup de mieux savoir ce que l'on demande à la détection (semi)automatique et de mieux formuler les critères utilisés pour le traitement informatique.

2.2 Détection (semi)-automatique

La détection semi-automatique s'applique essentiellement à des corpus écrits informatisés, qu'il s'agisse de documents papier informatisés comme un grand nombre de journaux maintenant, ou qu'il s'agisse de l'aspiration sur la toile de textes n'ayant pas d'autres modes d'existence (site, blog, etc.). Le volume et la diversité des textes qui peuvent être soumis à la détection de candidats néologiques sont énormes. Ce qui conduit à une réflexion sur les objectifs de la veille néologique et sur la qualité des néologismes, en fonction de leurs auteurs ou réémetteurs. On ne peut traiter de la même manière des lapsus calami et des créations délibérées et conscientes par exemple. Pour simplifier, deux grands types d'objectifs peuvent être distingués : celui qui consiste à mesurer l'évolution du lexique d'une langue et à lister les nouvelles unités qui circulent et qui ont vocation à être intégrés dans des dictionnaires de langue générale, et celui qui consiste à s'intéresser aux formes de la créativité lexicale des locuteurs d'une époque en fonction de leur statut, des domaines dont ils traitent, des conditions d'énonciation dans lesquelles ils profèrent leur énoncé, etc. (c'est ce choix qui a été fait dans Quemada 1993). Ce deuxième ensemble est beaucoup plus vaste que le premier car il prend en compte les hapax et les néologismes de diffusion restreinte. Mais il est intéressant de constater une moindre censure des Français face à la création lexicale que par le passé ainsi que les variations dans le type de matrices lexicales utilisées ou

dans les affixes mis en œuvre, etc. Les corpus journalistiques -du moins certains titres- permettent en revanche de voir ce qui circule et qui a des chances de s'intégrer à plus ou moins longue échéance. Ainsi les prêts « capés » ont été trouvés dans plusieurs quotidiens nationaux récemment (entre autres *Libération* et *Le Monde*), à l'occasion des conséquences de la crise des subprimes aux USA. Le mot, qui circule depuis plusieurs années dans les institutions financières, était encore absent du *Nouveau Littré* 2006, du *Petit Larousse* 2008 et il apparaît seulement dans le supplément « Mots nouveaux » du *Dictionnaire Hachette* 2009. Ceci témoigne de la diffusion récente de ce mot et de son entrée en train de se faire dans la langue générale[2].

Le traitement de corpus nombreux et divers génère beaucoup de bruit, et il est opportun de mettre en place des filtres destinés à le diminuer afin de décharger les linguistes de tâches manuelles de vérification et de tri. Une étudiante du master pro TILDE[3] a ainsi mis au point un certain nombre de filtres destinés à réduire le bruit de l'incorporation automatique et diminuer le nombre des candidats néologismes : des dictionnaires de toponymes, d'anthroponymes, de prénoms, des suites de trois lettres identiques ou des suites de lettres ne pouvant être dues qu'à des fautes de frappe, des problèmes de fausses coupes par agglutination ou déglutination, etc.

2.3. Recours à des moteurs de recherches

Les moteurs de recherche sont des outils intéressants pour procéder à la vérification de la diffusion d'un mot en général et d'un néologisme en particulier. Le nombre de sites répertoriés, la nature de ces sites, sont un premier indice de la diffusion d'un néologisme. La première information fera l'objet d'un champ dans la base de données Neologia (au moment de la constitution de la fiche). Des recherches plus fines peuvent être, au cas pas cas, conduites pour chercher les dates des occurrences mais elles ne sont pas toujours faciles à trouver. Ces premiers indices peuvent conduire à faire des enquêtes sur tel ou tel néologisme qui semble intéressant à étudier, de différents points de vue : taux de fréquence ascendant, descendant, pics, etc. ou bien type et nature des supports, etc. Les moteurs de recherche et les sites qu'ils indiquent fournissent ainsi des matériaux à traiter pour des mots clés révélateurs de l'évolution de la langue, de la société, etc. Ainsi l'emprunt *binge drinking* et ses équivalents français *cuite express* ou *biture express* sont-ils apparus au début 2007, dans des magazines et journaux, alors que les faits ainsi nommés existent depuis bien plus longtemps.

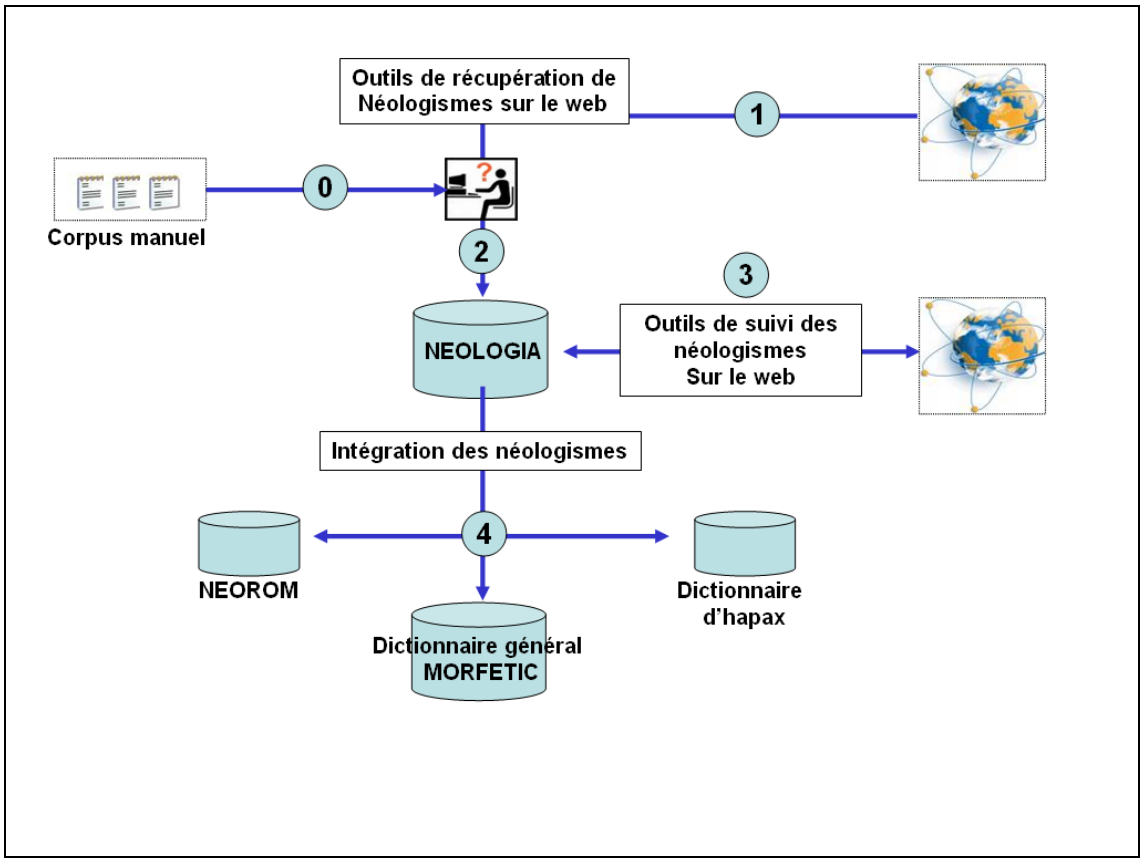
3. Exploitation des données : Neologia et les éléments d'une plateforme de veille néologique

Une fois les néologismes repérés, il reste à les stocker, à les analyser et à faire des études d'ensemble. C'est dans cette triple perspective qu'a été élaborée, en 2007, la base de données Neologia. Cette base de données prend place dans un système plus vaste de veille néologique, dont nous présentons tout d'abord l'architecture, avant de focaliser sur la base de données elle-même, qui en est l'élément le plus avancé.

3.1. Architecture générale d'un système de veille néologique

[2] Sa pénétration dans la langue courante, avec ses emplois multiples dans la presse généraliste fin 2007, assortis de gloses explicatives le fait changer de statut : ce n'est plus seulement un terme connu des seuls professionnels du domaine de la finance ou de ceux qui contractent des emprunts. Ce changement de statut relève de la néologie.

[3] MASTER PRO TILDE : le Master professionnel « Traitement Informatique et Linguistique des Documents Ecrits » est dispensé à l'université PARIS XIII depuis 2005.



Cette architecture comprend quatre grands processus indiqués dans les cercles.

Les processus 0 et 1 portent sur les sources documentaires utilisées pour constituer les bases de néologismes. Le processus 0, actuellement utilisé, consiste à repérer humainement les néologismes dans les discours écrits et oraux. Le processus 1, en cours de développement, permettra de récupérer automatiquement sur le web des néologismes à partir d'une analyse morpho-syntaxique des corpus numériques (voir 3.3 pour plus de détails). Le processus 2 est le processus de saisie (et de validation pour les candidats néologismes) des néologismes dans la base de donnée Neologia.

Le processus 3 permettra de suivre l'évolution du néologisme dans les discours, en informant sur la circulation des mots sur le web, diachroniquement et synchroniquement (en repérant par exemple que tel néologisme se cantonne à tel domaine, ou commence à s'étendre à tel autre etc.). Ce processus est également en cours de réalisation.

Le processus 4 permet de reverser les néologismes dans d'autres dictionnaires : en premier lieu, intégration dans un dictionnaire général, lorsqu'il sera estimé que le terme n'est plus un néologisme ; en second lieu, intégration dans un dictionnaire des hapax ou des néologismes de très faible diffusion, si l'usage du terme s'avère ponctuel et que sa courbe de fréquence est nulle ou basse. À noter que ces deux processus vont tendre à l'automatisation avec l'utilisation du module 3 de suivi de l'évolution du terme sur le web. Nous avons également noté ici l'exportation de certaines de données de Neologia dans la base Neorom, destinée à comparer la néologie des langues romanes[4].

3.2. Fonctionnalités de la base Neologia

La base Neologia, après une période de rodage, est entrée dans une phase de révision avant sa montée en puissance au cours de l'année 2008. Bâtie par Emmanuel Cartier, elle offre un certain nombre de fonctionnalités destinées à remplir la grille d'analyse élaborée et remaniée régulièrement par J.F. Sablayrolles depuis une vingtaine d'années. Elle permet ainsi des requêtes multiples, simples ou croisées et elle entretient des rapports interactifs avec deux autres outils extérieurs.

[4] Y coopèrent des équipes relevant des néologismes catalans, espagnols, galiciens, italiens, roumains, portugais, français de France et du Québec…, à l'initiative de Térésa Cabré de l'Université Pompeu Fabra de Barcelone.

Pour faciliter l'accès des utilisateurs à la base Neologia, celle-ci a été installée sur un serveur web. Elle est donc accessible via n'importe quel navigateur web, quels que soient la situation géographique et l'environnement informatique des utilisateurs.

Trois niveaux d'utilisateurs ont été définis : administrateur, auteur et invité. Les administrateurs gèrent la structure de la base, ont la possibilité de valider les entrées saisies par les auteurs. Ils disposent également de tous les droits propres aux auteurs : ces derniers peuvent saisir de nouvelles entrées, les modifier et les supprimer, ainsi qu'ajouter de nouveaux contextes à l'ensemble des entrées de la base. Un statut invité est enfin prévu pour des collègues extérieurs qui souhaiteraient consulter la base des néologismes validés par les administrateurs. Sous certaines conditions, un mot de passe leur serait fourni pour qu'ils puissent visualiser la base.

L'accès aux entrées du dictionnaire se fait via un moteur de recherche permettant de filtrer les entrées selon l'un ou plusieurs de tous les champs utilisés pour décrire une entrée. Par défaut, le moteur de recherche ne comporte aucun filtre, et l'ensemble des entrées est présenté sous forme de tableau synthétique triable. L'interface se compose d'une zone « recherche » et d'une zone « résultat ». La zone « recherche » permet aux utilisateurs de sélectionner un ou plusieurs des champs renseignés pour chaque terme. La zone « résultat » présente, sous forme de tableau, les différents termes répondant aux critères de la recherche. Deux boutons, à gauche de chaque ligne, permettent l'un de visualiser la fiche complète du terme, et l'autre de la supprimer. La fiche complète est divisée en deux zones : une zone principale, en haut, qui présente les informations génériques liées à l'item et une zone secondaire, sous forme d'onglets, qui permet d'accéder à différentes composantes descriptives.

Comme la phase de rodage en a montré l'utilité, cette base peut évoluer, tant dans le nombre des champs descriptifs (par ajout ou suppression) que dans les valeurs prévues dans les champs qui ont des menus déroulants fermés. On a ainsi fondu en un seul champ « guillemets » les deux valeurs « guillemets simples » et « guillemets doubles » dont la distinction ne s'avère pas pertinente. En revanche, nous introduirons deux nouveaux champs dans la partie contexte afin de distinguer systématiquement le créateur du néologisme (et son statut : journaliste, etc.) et son énonciateur / (ré-)émetteur (et son statut). Les deux peuvent coïncider, mais c'est loin d'être toujours le cas et le mode de signalement de cette distinction s'est avéré insuffisant dans la première version de la base. Dans l'exemple reproduit ci-dessous, en 3.3.6, l'identité de l'énonciateur et son statut sont encore indiqués dans le champ commentaire du contexte.

3.2. Deux liens avec l'extérieur

La base Neologia entretient deux liens privilégiés avec le monde extérieur : d'une part avec les autres dictionnaires élaborés au laboratoire LDI, d'autre part avec un extracteur de néologismes formels en cours de mise au point dans ce même laboratoire.

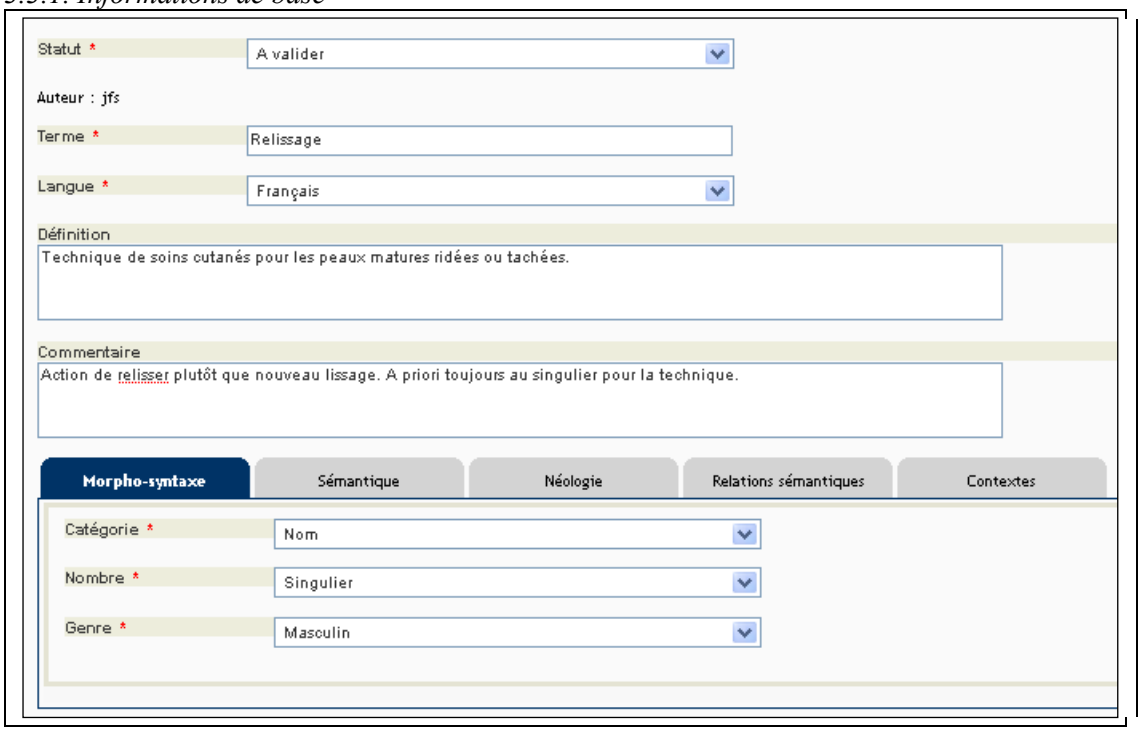
Les dictionnaires élaborées au LDI, dont Morfetik, mentionné en début d'article, accueilleront les néologismes qui se sont diffusés au point de perdre leur caractère néologique. L'exportation de ces unités lexicales est assurée par une fonctionnalité informatique et facilitée par la similitude des types de description en vigueur au sein du laboratoire. Non seulement ces dictionnaires devront accueillir de nouvelles entrées, mais il est inévitable aussi que des entrées anciennes doivent être revues du fait de modifications introduites par l'apparition d'un nouvel élément dans la classe d'objets ou le champ sémantique auxquels elles appartiennent.

Un outil d'extraction automatique de néologismes formels alimentera également la base en sus des collectes manuelles. Télanaute, réalisation de Fabrice Issac, arpente des sites sur la toile et utilise Morfetik comme corpus d'exclusion pour livrer des candidats néologismes. Il y a encore beaucoup de bruit que l'on s'efforce de diminuer avec des modifications dans les requêtes et par l'ajout de filtres auxquels on ajoute régulièrement de nouvelles unités. Ce qui reste est soit entré dans le dictionnaire Morfetik s'il s'avère qu'il s'agit d'une lacune d'une unité non néologique, soit entré dans la base Neologia.

3.3. Une fiche Neologia : relissage

Outre les premières informations de base : statut administratif de la fiche (à valider / validé), la vedette (sous forme lemmatisée sauf exceptions motivées), la langue (actuellement seulement le français), une définition de type lexicographique et un champ commentaire libre (pour toutes les informations jugées nécessaires par le créateur de la fiche et qui n'entrent pas dans la grille d'analyse prévue), une fiche comprend cinq onglets correspondant à cinq ensembles d'informations linguistiques : morpho-syntaxique, syntactico-sémantique, propriétés néologiques, relations sémantiques et contextes.

3.3.1. Informations de base

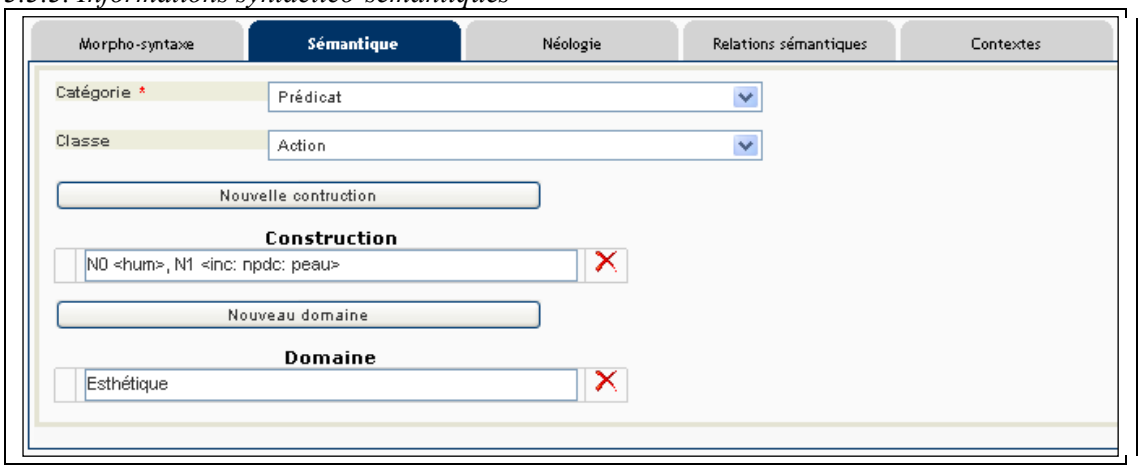


Ecran 1 : vue globale d'une fiche NEOLOGIA : informations de base et morpho-syntaxiques

3.3.2. Informations morpho-sémantiques

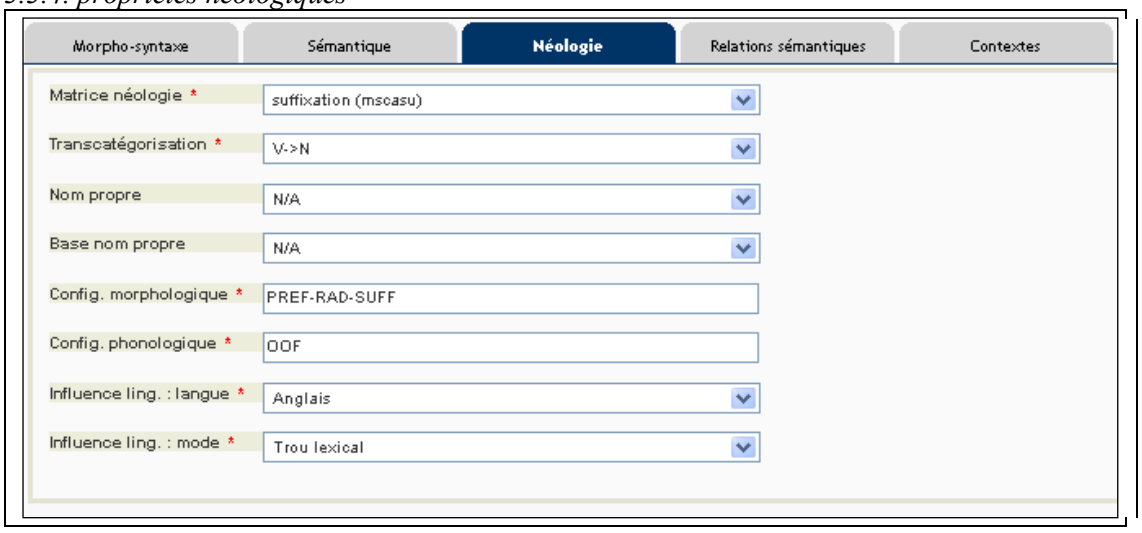
Les informations concernent ici la partie du discours, le modèle de flexion, et diverses informations concernant les catégories grammaticales de l'item. Toutes ces informations permettront l'intégration éventuelle des néologismes dans la base générale Morfetik.

3.3.3. Informations syntactico-sémantiques



Cet onglet présente des informations sémantiques obéissant au modèle développé au LDI : catégorie syntactico-sémantique du néologisme (prédicat, argument ou actualisateur), classe sémantique du terme, schéma(s) syntactico-sémantique(s) pour les prédicats, domaine(s) du terme. Ainsi *relissage* est-il un prédicat d'action à deux arguments, le premier peut être défini par le seul trait humain, le second appartient à la classe d'objet des inanimés concrets (inc), noms de parties du corps (npdc) et plus précisément la peau. Ce néologisme relève du domaine de l'esthétique. Là encore, ces informations permettront d'intégrer les néologismes dans la base lexicographique développée au LDI.

3.3.4. propriétés néologiques



Ce niveau, propre aux néologismes, permet d'abord d'indiquer la matrice lexicale responsable de la création du néologisme. La liste des matrices est inspirée de celle de J. Tournier (1985) revue par Jean-François Sablayrolles (2000) et encore légèrement améliorée depuis (Pruvost et Sablayrolles 2003, Benhariz à paraître en 2009). Ce n'est pas un simple catalogue mais un tableau fortement hiérarchisé (à cinq niveaux) qui regroupe une vingtaine de matrices en ensembles, sous-ensembles, sous sous-ensembles… sur la base de propriétés communes. Les codes associés aux matrices (comme mscasu pour « suffixation » qui équivaut à morpho-sémantique, par construction, par affixation suffixale) permettent de faire des calculs pour tous les regroupements par niveaux possibles. Une requête msca (matrice interne, morpho-sémantique, par construction, par affixation), par exemple donne le sous-ensemble formé par la préfixation, la suffixation, la dérivation inverse et la construction parasynthétique.

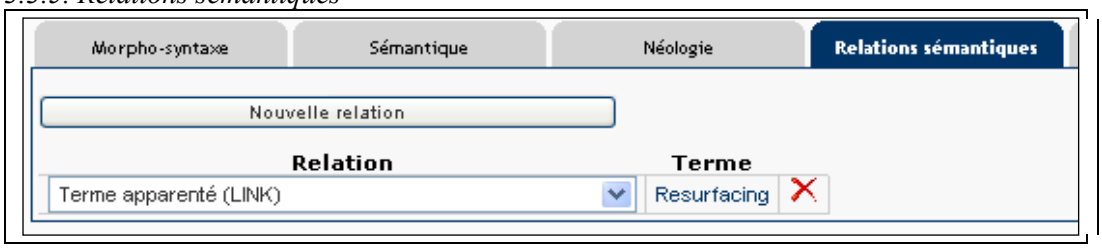
Le champ mode d'influence linguistique est destiné à noter les formations françaises (dérivées, composées…) dues à l'existence de mots étrangers sans qu'il s'agisse à proprement parler d'emprunts : calques, faux-emprunts, équivalents (pour combler un trou lexical en français comme *relissage* équivalent de *resurfacing*), structure (les séries de composés régressifs comme *royale attitude*… ou *serial menteur*…) …

Beaucoup de néologismes étant l'adaptation au cotexte d'un mot appartenant à une partie du discours autre que celle requise par ce cotexte, on note sous l'étiquette générale « transcatégorisation » tous les changements de catégories grammaticales entre la base et le néologisme, quelle que soit la matrice lexicale qui opère cette transcatégorisation (préfixation, suffixation dérivation inverse, conversion, déflexivation…).

Deux champs sont réservés à la catégorie nom propre, le premier pour la création de noms propres (qui ne seront pas intégrés dans les dictionnaires généraux sauf si, par antonomase, ils deviennent noms communs) et le second pour les néologismes formés sur la base de noms propres. Dans chacun de ces deux champs, on distingue des anthroponymes, des toponymes et des noms de marques, avec A, T, M pour le premier champ et a, t, m pour le second)

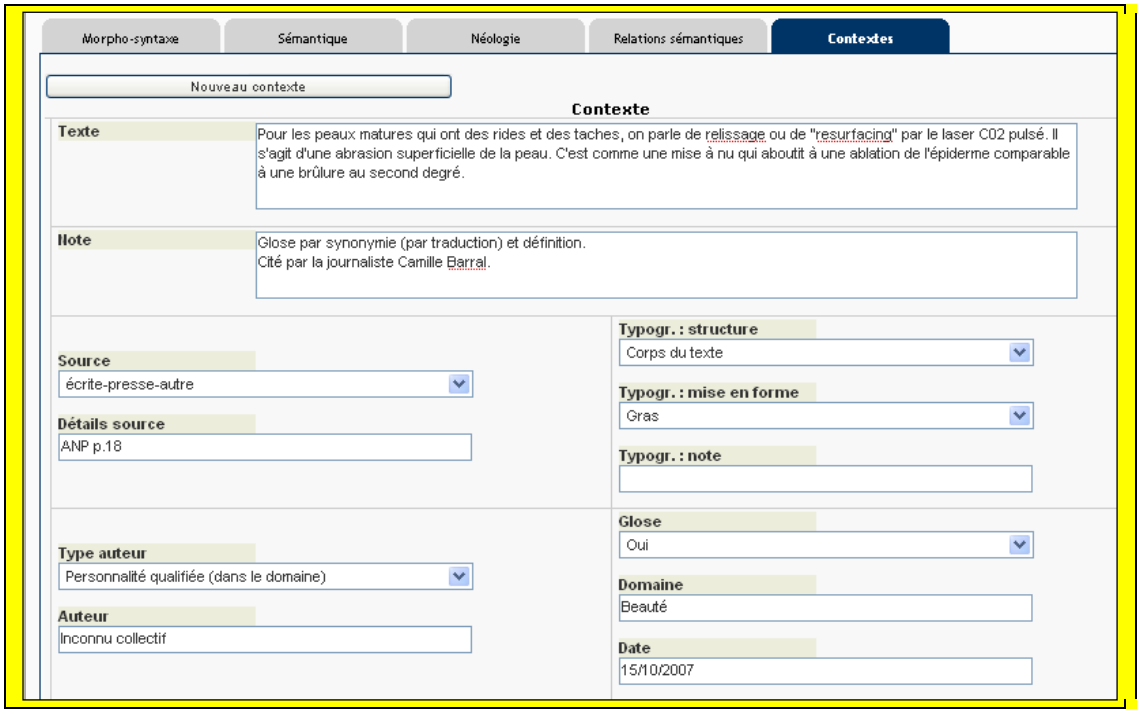
Deux informations non spécifiques aux néologismes concernent leur structure morphologique et syllabique (O pour syllabe ouverte et F pour syllabe fermée, du point de vue phonétique). L'objectif est double. D'une part se donner la possibilité de comparer la « physionomie » des néologismes par rapport aux mots du lexique conventionnel (longueur, complexité, structure…) : elle se distingue du « mot français prototypique » constitué de deux syllabes (ouvertes le plus souvent) si l'on en croit Dauzat et des études statistiques. D'autre part, le premier de ces deux champs permet de bien distinguer la matrice lexicale qui produit le néologisme et l'analyse morphologique de celui-ci : la forme verbale *désagrémente* est constitué d'un préfixe (*dés-*), d'un radical (*agré*), d'un suffixe (*-ment*) et d'une marque flexionnelle graphique (*-e*). Créé par préfixation, il signifierait 'ôter une amélioration' d'où 'enlaidir' mais le contexte où il a été émis *Quelle affaire te désagrémente ?* conduit à y voir une conversion du nom *désagrément* en verbe signifiant 'causer une sensation déplaisante'. Pour des raisons sémantiques également, *relissage*, à la fois préfixé et suffixé, est plutôt créé par suffixation sur le verbe *relisser* que par préfixation sur le nom *lissage*. On a ainsi une chaîne dérivative : *lisser -> relisser -> relissage* plutôt que *lisser -> lissage -> relissage*.

3.3.5. Relations sémantiques



Cet onglet permet d'expliciter un certain nombre de relations sémantiques (synonymie, hyper- et hyponymie, relations partie-tout, etc.) entre le terme et d'autres termes de la langue concernée ou une relation de traduction dans des cas de créations produites pour combler un trou lexical.

3.3.6. Contextes



Ce pan descriptif permet de citer le ou les contextes dans le/lesquels les néologismes ont été repérés[5]. Plusieurs contextes peuvent en effet être indiqués, en particulier pour des attestations antérieures à celle qui a fait entrer le néologisme dans la base ou pour des contextes particulièrement éclairants ou significatifs. Des informations très fines permettent de catégoriser les différents contextes, leurs paramètres spatio-temporels, les créateurs et les (ré)émetteurs (identité et type). Dans le cas où le créateur du néologisme ne coïncide pas avec l'auteur du texte dans lequel il est cité ou simplement utilisé, on note l'identité et le statut de ce dernier dans le champ commentaire (en attendant la création de deux champs supplémentaires), réservant les champs auteur et type auteur aux créateurs, connus ou non, des néologismes… Le nombre d'attestations relevées à un moment donné à diverses sources (dont les moteurs de recherche) sera éventuellement indiqué dans le champ commentaire, en attendant la mise en service des systèmes automatiques indiquées antérieurement. Les caractéristiques typographiques (titre, chapeau, corps du texte, légende, bulle… ; guillemets, italiques, gras, couleur différente…) ainsi que d'éventuelles gloses sont également systématiquement indiquées.

3.4. Exemples de requête

Pour illustrer l'intérêt de cette base et sa souplesse, un exemple, en plusieurs étapes, suffira. Une première requête portant sur un seul critère, la catégorie grammaticale, permet de faire apparaître les 76 verbes néologiques, représentant un peu moins de 10 % des 797 néologismes actuellement entrés dans la base. Une deuxième requête ajoutant la présence d'une influence anglaise montre que plus d'un quart de ces verbes

[5] ANP désigne l'hebdomadaire gratuit diffusé dans le métro *À Nous Paris*.

néologiques sont concernés : 20 sur 76. Un troisième ensemble de requêtes actionnant le filtre « matrice » donne 6 emprunts véritables (*breaker* « percer, pour un comédien », *gatecrasher*, *pumper*, *spoofer*, *trianguler* « s'emparer des valeurs de son adversaire politique », *winner*), 5 conversions à partir de noms empruntés (*facebooker*, *groover*, *hightecher*, *loller*, *nexter*), 2 suffixations (*se conceptoriser*, *pipoliser*), 2 cas de changement de sens de mots existants (*commissionner* « commanditer » et *signer* « embaucher »), une préfixation (*détrianguler*), un mot-valise (*E-naugurer* « inaugurer un site informatique »), une dérivation inverse (*brainstormer*), un détournement (*dognapper*), une nouvelle combinatoire syntaxique (la construction intransitive de *flasher* « faire de l'effet, sortir de l'ordinaire »).

D'autres requêtes pourraient encore être ajoutées à celles-ci. C'est, en fait, une multiplicité des types et des combinaisons de requêtes qui est possible. Le problème sera de sélectionner les associations de critères pertinentes, c'est-à-dire celles qui sont révélatrices de la créativité lexicale du français contemporain et de l'évolution de son lexique.

Conclusion

Cet article a présenté deux interactions majeures entre la néologie et l'informatique : la veille néologique et les bases de données. L'outil central est une base de données, Neologia, permettant de stocker les néologismes. L'interface qui a été construite est maintenant en état de fonctionnement, et la phase d'expérimentation est bien avancée. Outre la présentation de cette base, ont été décrites deux interactions majeures qu'elle doit avoir avec le monde extérieur : lien avec les corpus extérieurs pour récupérer des candidats néologismes et liens avec un dictionnaire général pour intégrer les néologismes passés dans l'usage. Cependant, beaucoup de travail reste à faire car le matériau se renouvelle sans cesse : la langue génère sans cesse de nouveaux termes, et il s'agit véritablement de se mettre totalement à l'écoute des corpus produits, en connectant la base aux corpus au moment où ils deviennent disponibles. L'informatisation de données, le recours à des outils informatiques pour les incorporations et les analyses permettent de traiter des données incomparablement plus nombreuses qu'auparavant et de faire des recherches qui n'étaient pas possibles.

Références bibliographiques

BENHARIZ-OUENNICHE Soundous (à paraître), « Diminuer les fluctuations du sentiment néologique » *Neologica* n° 3, 2009.

CARTIER Emmanuel (2007), « Intégration des prédicats verbaux dans l'analyseur sémantique TextBox : l'exemple des verbes de cognition ». Dans Grezka Aude, Martin-Berthet Françoise, (éd), 2007, Verbes et classes sémantiques, *Verbum* 2007-1, Presses universitaires de Nancy, p. 97-112.

CARTIER Emmanuel (2007), « TextBox, a Written Corpus Tool for Linguistic Analysis ». In Fairon Cédrick, Naets Hubert, Kilgarriff Adam, De Schryver Gilles-Maurice, (éd), Building and Exploring Web Corpora (WAC3 - 2007), *Cahiers du CENTAL* 4, p. 33-42, Presses universitaires de Louvain, Louvain-la-Neuve.

CARTIER Emmanuel (2008), « Reconnaissance automatique des séquences figées », Journée Conscila « Figement », 15 février 2008, *Zeitschrift für französische Sprache und Literatur*.

CORBIN Danielle (1990), « Homonymie structurelle et définition des mots construits, vers un dictionnaire dérivationnel », *La définition*, J. Chaurand et F. Mazière éd., Larousse, pp. 175-192.

DAUZAT Albert (1943), *Le génie de la langue française*, Payot.

GARDIN B., LEFEVRE G., MARCELLESI C., MORTUREUX M.-F. (1974), « À propos du sentiment néologique », *Langages* n° 36, p. 45-52.

GREZKA Aude, MARTIN-BERTHET Françoise (éd) (2007), « Présentation », Verbes et classes sémantiques, *Verbum* 2007-1, Presses universitaires de Nancy, p. 3-10.

GROSS Gaston (1994), « Classes d'objets et description des verbes », *Langages*, 115, p. 15-30.

MATHIEU-COLAS Michel (2006), « Dictionnaire morphologique du français, I. Formes simples (1996-2006) », Rapport technique du LLI, Villetaneuse, Université Paris 13.

PRUVOST Jean et SABLAYROLLES Jean-François (2003), *Les néologismes*, Que sais-je ?, PUF.

QUEMADA Bernard (1993), (sous la direction de), *Mots nouveaux contemporains 1*, Matériaux pour l'histoire du vocabulaire français, CNRS, Paris, Klincksieck.

REY Alain (1976), « Néologisme, un pseudo concept ? », *Cahiers de lexicologie* n° 28, pp. 3-17.

SABLAYROLLES Jean-François (2000), *La néologie en français contemporain* " examen du concept et analyse de productions néologiques récentes ", coll. Lexica Mots et Dictionnaires, Champion.

SABLAYROLLES Jean-François (2002), « Fondements théoriques des difficultés pratiques du traitement des néologismes », *Revue française de linguistique appliquée*, vol. VII-1. / juin 2002 « Lexique : recherches actuelles », p. 97-111.

SABLAYROLLES Jean-François (2003), « Le sentiment néologique », *L'innovation lexicale*, J.-F. Sablayrolles éd., Champion, pp. 279-295.
SABLAYROLLES Jean-François (2008), « Néologie et dictionnaire(s) comme corpus d'exclusion », *Néologie et terminologie dans la lexicographie francophone*, J.-F. Sablayrolles éd., coll. Lexica, Champion p. 19-36.
SABLAYROLLES Jean-François (à paraître), « Le militantisme néologique dans sept dictionnaires du XIX^{e} siècle », *De l'engagement dans les dictionnaires*, version enrichie des actes du colloque « La lexicographie militante », Jussieu, 8 et 9 décembre 2006, François Gaudin (éd).
TOURATIER Christian (2002), *Morphologie et morphématique, Analyse en morphèmes*, Langue et langage n°8, Publications de l'Université de Provence, 2002.
TOURNIER Jean (1985), *Introduction descriptive à la lexicogénétique de l'anglais contemporain*, Paris-Genève, Champion-Slatkine.